EIEPHãT: Early English Print in the HathiTrust
a Linked Semantic Worksets Prototype

Final report
Principal Investigator: Kevin Page, Oxford e-Research Centre, University of Oxford
Co-Investigator: Pip Willcox, Bodleian Libraries, University of Oxford

The EIEPHãT project – Early English Print in HathiTrust, a Linked Semantic Worksets Prototype – demonstrates the use of Linked Data for combining, through worksets, information from independent collections into a coherent view which can be studied and analyzed to facilitate and improve academic investigation of the constituents.

1. Background

Early English Books Online Text Creation Partnership (EEBO-TCP) is a partnership with ProQuest and more than 150 libraries and universities, led by the universities of Michigan and Oxford, to generate highly accurate, fully-searchable, XML-encoded texts corresponding to books from ProQuest’s Early English Books Online database.

The EEBO corpus consists of the works represented in the English Short Title Catalogue (STC) I and II (based on the Pollard & Redgrave, and Wing short title catalogues), as well as the Thomason Tracts and the Early English Books Tract Supplement. Together these trace the history of English thought and learning from the first book printed in English in 1473 through to 1700. From 2000-2009, EEBO-TCP Phase I successfully converted 25,000 selected texts from the EEBO corpus into TEI-compliant, XML-encoded hand-transcribed texts, which were made freely available to the public in January 2015.

2. Project objectives

The EIEPHãT project focuses on the potential symbiosis between two datasets as an illustration of how the Semantic Web (including RDF, SPARQL and ontologies) can provide a technological foundation for applications grounded in scholarly investigation (figure 1). The first dataset is EEBO-TCP, as described above; and the second is a custom HathiTrust dataset consisting of all materials described in their metadata as being in English and published between 1470 and 1700, with variable bibliographic metadata inherited from the contributing institutions.

The key objectives of EIEPHãT were to generate an RDF export of metadata extracted from EEBO-TCP which, when combined with RDF from the HathiTrust Research Center using appropriate ontologies, could be employed to create prototype interfaces through which a scholar might perform seamless investigations over the combined linked dataset.

Figure 1. EIEPHãT overview
3. Project personnel

The EIEPHâT project commenced in June 2014 and was completed in April 2015. The principal investigator of the project was Dr Kevin Page of the University of Oxford e-Research Centre, with Pip Willcox, Curator of Digital Special Collections at the Bodleian Libraries, as his co-investigator. Research effort was concentrated within the Oxford e-Research Centre, with Dr Terhi Nurmikko-Fuller joining the team in September 2014 as the primary researcher; John Pybus, a senior researcher in the Centre, contributed to initial requirements analyses, and valuable input was made by David Weigl during the latter stages of the project and during deployment of the SALT software (section 4.4), of which he is the core developer.

The project team reported to a Technical and Scholarly Steering Committee for advice and guidance, which met 14 times over the duration of the project (approximately fortnightly) and was formed of experienced members of the e-Research Centre and Bodleian Libraries: Professor David De Roure (Director of Oxford e-Research Centre), and the Bodleian Libraries’ Michael Popham (Head of Digital Collections and Preservation Services), Dr Christine Madsen (Head of Digital Programmes), and Judith Siefring (Digital Editor).

4. Project activities and accomplishments

The project has undertaken a number of research activities, often in parallel, in pursuit of the above objectives. We report on these outcomes in the following subsections.

4.1 Consultative workshop: validation of workset requirements and utility

On 12 June 2014, the EIEPHâT team organized a consultative workshop with academics and other potential audience members for the tools created by the project. From the outset we had planned that scoping and prioritization of the project’s technical developments were to be informed by actual scholarly questions guided by the community, and iterated in consultation with them.

To this end, we invited 12 colleagues from 5 universities (Bath Spa, London, Oxford, Oxford Brookes, and Reading), two libraries (the Bodleian and the British Library), and a publisher (Oxford University Press) to the Oxford e-Research Centre for a day of discussion, formal and informal. As well as our Technical and Scholarly Steering Committee members, librarians, digital curators, and publishers, our guests included academics across career stages, from students to professors.

EIEPHâT was put into its wider context before the group broke out into smaller discussion groups to consider the questions we had raised about how the tools might best work for them. The groups also considered the sample academic questions outlined in the project proposal, those highlighted in the SECT\(^1\) project, and came up with research questions of their own that were refined during the closing plenary discussion. Attendees also filled in a questionnaire, to give us quantitative data to guide our decisions. The day’s discussions influenced softer aspects of the project as well as its motivating technical questions, for example the need for an interface for refining worksets that humanists would be at home using. Several of the academics have kept in close touch with the project. Two (from Bath Spa and Reading) returned for a consultation with Dr. Nurmikko-Fuller in the autumn. One of these is now incorporating more digital humanities and curation training in his undergraduate teaching. He also came to the EEBO-TCP hackfest (March 2015) and was the academic lead on one of the prize-winning teams.

---

1. [http://blogs.bodleian.ox.ac.uk/eebotcp/files/2012/05/SECTWorkshopSummaryReport.pdf](http://blogs.bodleian.ox.ac.uk/eebotcp/files/2012/05/SECTWorkshopSummaryReport.pdf)
A more detailed report of the workshop can be found as an appendix to this report.

4.2 Technical analysis of metadata requirements and ontology selection

In light of the domain motivations articulated at the scholars’ workshop, a technical analysis of the metadata requirements to support scholarly worksets bridging the two datasets was undertaken. In collaboration with the WCSA team, we surveyed both the addressable resources and the schema expressivity of ontologies that could potentially be used to parameterise these types of workset, including MODS RDF, Bibframe, Schema.org and FRBRoo. A full description of this work can be found in Nurmikko-Fuller et al.²

We then identified parsable information structures and terms in the EEBO-TCP Text Encoding Initiative (TEI) data which were suitable for parameterizing worksets and, informed by the survey, selected ontology elements used to encode this EEBO-TCP metadata, which would be compatible (or at least, for our purposes, comparable) with the RDF structures being generated for HathiTrust metadata by the WCSA team. The resultant ontology collection - the Early English Books Online Ontology, or EEBOO - includes selections from MODS, Bibframe, and PROV, along with custom elements to encode additional structures (e.g. dates); and in addition encodes worksets using the WCSA workset ontology or, alternatively, the Research Object (RO) model. EEBOO can be found in the project github repository.

4.3 Tools for generation of custom RDF derived from EEBO-TCP

Several tools and scripts were re-used or developed to generate RDF triples, conforming to EEBOO, from the derived from EEBO-TCP TEI headers. These tools and configurations can be found in the project github repository.

First a set of python scripts was developed to process the TEI P5 XML, then the Karma Data Integration Tool³, from the University of California Information Sciences Institute, was used to map the EEBO-TCP data structures into the EEBOO ontology and output RDF. Particular attention was paid to dates encoded within strings in EEBO-TCP, as an illustration of rich semi-structured data that can be extracted into structured RDF (there are numerous other similar metadata elements that could form the basis of extension work). These were matched using regular expression rules and parsed for encoding using corresponding temporal extensions to EEBOO. We matched 98 distinct expressions of author date using 75 regular expression, and 1,510 distinct expressions of publication date, parsed into 54 types. Finally RDF links to author records in VIAF and the Library of Congress, and multimedia pages for texts in JISC Digital Books and HathiTrust, were generated and added to the EEBOO triples.

We took 24,926 EEBO-TCP Phase 1 records with 22 distinct types of information in the headers, including 6 different ID numbers and 3 mains types of date (publication date of historical work, author associated historical date(s), XML publication/editing dates). EEBOO incorporates 7 of these datatypes, and extends into subcategories for author names and different types of date. After processing the EEBO-TCP headers, EEBOO contains 713 unique places, 6,489 unique expressions of Person of which 3,588 have VIAF and LoC IDs.

³ http://www.isi.edu/integration/karma/
4.4 Alignment of instance data between EEBOO and HathiTrust using ‘SALT’

The Semantic Alignment and Linking Tool (SALT) was developed at the Oxford e-Research Centre to simplify the process of confirming alignments between datasets, prioritising potential matches to bring them to the attention of the scholar or expert by using both string and semantic-contextual distances. SALT is designed on the basis of assisting, not completely automating, the process of making a scholarly judgement during alignment. SALT accesses data held in RDF triplestores using a semantic configuration for the (potentially multiple) schemas used; it then writes alignments and match decisions back to a triplestore enabling layering of assertions and generation of provenance. SALT was configured to access both the EEBOO and HathiTrust custom dataset RDF, then used to align author names between the two (Figure 2).

4.5 Provisioning an RDF triplestore populated with EEBOO and HathiTrust RDF

Having collected and generated the RDF triples as outlined in the previous sections, we populated these datasets into a Virtuoso RDF triplestore which could provide the necessary SPARQL query interface upon which the user interfaces (described in the following section) are built. The triplestore was initially configured on several development workstations, then migrated to a server for final demonstration, and ultimately contained named graphs for EEBOO, alignment and match decisions from SALT, a copy of the WCSA-produced graph of the HathiTrust custom data set, and worksets as generated by the prototype UIs. While the named graphs are all currently all co-located within one Virtuoso instance, the design of the tools in Section 4.6 is such that they could be transparently distributed across different hosts and institutions should this be desired.

The EiEPHâT Virtuoso triplestore contains 1,137,502 triples, consisting of: 251,725 entities, 66 distinct classes, 214 distinct predicates. 287,581 distinct subject nodes and 294,677 distinct objects. A breakdown by named graph yields: 468,022 EEBOO triples, 341,949 triples aligning EEBOO and HathiTrust, 309,257 triples from the WCSA produced HathiTrust custom dataset RDF, 11,168 SALT match decisions, and 1,354 triples for workset descriptions.

4.6 Implementation of prototype architecture and user interfaces for workset construction and viewing

The data extraction, ontological, and alignment outputs outlined in the previous sections enable us to make SPARQL queries to the triplestore that semantically match and return works from both EEBOO and the HathiTrust custom dataset, and which build provide the types of investigative construction requested at the

---


5 The SPARL interface is at: [http://eeboo.oerc.ox.ac.uk/sparql](http://eeboo.oerc.ox.ac.uk/sparql); please email kevin.page@oerc.ox.ac.uk for access.
consultative workshop. The following are paraphrases of SPARQL queries that will successfully return results from the EiEPHãT architecture and datasets:

- Find all the works, appearing in both datasets, which have been written by Richard Baxter.
- Find works in both datasets that have been published in Oxford.
- Find works which were published outside of London (where the bulk were published).
- Find all the works contained in the two datasets for authors who have at least once published on the subject of “Political science”.
- Find all the works contained in these two datasets for authors who have at least once published works which are categorised as “biography”.

SPARQL queries such as these provide the basis for our prototype user interfaces: a workset constructor that allows a scholar to specify a number of attributes by which a workset is assembled; and a workset viewer that enables the scholar to inspect the contents of a workset and follow links to further resources (e.g. the EEBO-TCP text or images).

The workset constructor (figure 3) defines a workset using parameters selected by the user which are then, in the background, assembled into the SPARQL queries used to create a workset. The interface is automatically populated with valid potential attributes that are themselves retrieved from the triplestore using SPARQL queries, using ontological terms asserted to have equivalent meaning between the datasets. The workset viewer (also figure 3) then retrieves RDF workset contents, record metadata, data links, and multimedia links (to the JISC Historic Books collection or the HathiTrust Digital Library). The underlying workset data can be retrieved according to the draft WCSA workset ontology, or using the Research Object model, and small ‘beans’ are used throughout the interface to illustrate where an element has an underlying EEBOO or workset RDF URI.

* It is of particular note that this returns results across both datasets since our EEBO-TCP import contained no genre or topic information; this associated is entirely inferred from the semantic links.
Both web applications are written in Python, using the Flask framework, and both rely heavily on the semantic information encoded in RDF and queried using SPARQL. Figure 4 shows the overall composition of the final ElEPHãT architecture. The online prototype\(^7\) has been populated with exemplar worksets based on the scholarly questions listed above.

**Summary**

The ElEPHãT prototype has demonstrated the utility of using RDF and Linked Data to bring together HathiTrust content with ‘boutique’ collections such as EEBO-TCP, providing a template for how this can be achieved. While the project has highlighted the potential complexity of generating structured metadata from new sources, we believe it also shows how RDF provides a common ‘middle ground’ that can decouple the process of aligning items, from the preparatory re-encoding (or restructuring) of external corpora that are to be incorporated or linked. ElEPHãT also highlights that, using RDF, this can be beneficially performed in a piecemeal or iterative fashion - an obvious future addition being the RDF encoding of other EEBO-TCP headers including remaining date fields and imprints. In linking between HathiTrust and EEBO-TCP we have also demonstrated how either – or both – of these might be further linked using tools such as SALT. In this sense, we have only scratched the surface of the information held here, not to mention the possibilities offered by the content of the texts; while ElEPHãT was conceived and implemented as a limited prototyping project, we believe the methods it has proved can, and should, be more widely deployed in both scope and ambition.

Kevin Page  
*University of Oxford e-Research Centre*  
*Oxford, 15th June 2015*

\(^7\) [http://eeboo.oerc.ox.ac.uk/](http://eeboo.oerc.ox.ac.uk/)
Dissemination

Publications


Talks

Making Links: Connecting humanities resources for scholarly investigation, International Conference on Digital Humanities in Japan, JADH2015, Kyoto University, September 2015.

Trunk to tail: linking ElEPHãTs through the Semantic Web, Centre for Digital Scholarship seminar, University of Oxford, June 2015.

ElEPHãT: Early English Print in HathiTrust, a Linked Semantic Worksets Prototype, HathiTrust Research Center Uncamp, University of Michigan, March 2015.


Exercising ourselves “in the Analysis of many examples”: corpora, collaboration, and communication, a digital humanities masterclass at the Institute of Advanced Studies, University of Western Australia, August 2014

Outreach and teaching

ElEPHãT consultative workshop, Oxford, June 2014 (see section 4.1 and attached report).

The ElEPHãT team attended the EEBO-TCO Hackfast at the Bodleian Weston Library in March 2015, and made the EEBOO triplestore available to participants. During the event we produced an initial alignment between EEBOO and the Oxford University Press ‘Early English Authorities’ data.

At the Digital Humanities at Oxford Summer School in July 2015, ElEPHãT will be disseminated in two workshops (Digital Approaches Medieval and Renaissance studies; and Humanities Data: Curation, Analysis, Access, and Reuse) and will form the basis of examples throughout the week in a third workshop, on Linked Data.
The Oxford Illinois Digital Libraries Placement Programme (OIDLPP) will host a student in the Oxford e-Research Centre between July and August 2015, who will work on a project extending the ELPHat data to include information from EEBO-TCP imprint fields.

**Deliverables**

1. Report on consultative workshop (attached as an appendix).
2. Software to produce RDF from EEBO-TCP and support alignment with HathiTrust and external entities (code to be uploaded to github repository). Including Python scripts, regular expressions, Karma documentation/configuration (as described in section 4.3).
3. Documentation/configuration of SALT for use with EEBOO and the HathiTrust custom dataset (to uploaded to the github repository; as described in section 4.4).
4. Software implementing the workset constructor and viewer interfaces (to be uploaded to the project github repository; as described in section 4.6).
5. This final project report.

The github code repositories for ELPHat will be populated during summer 2015 under the project space at: [https://github.com/oerc-ellephat](https://github.com/oerc-ellephat)

**Explanation for variance of costs**

The attached financial summary shows an overspend in travel expenses compared to the original, relatively small, travel budget. This is due to payment of expenses towards presenting the accepted JCDL paper (see Publications, above) at conference; and of underestimated costs for travelling to the final WCSA reporting meeting (this was originally budgeted for a single day event in Chicago; additional costs were incurred due to length and location of the 3 day event in Ann Arbor including the HathiTrust UnCamp, and coincidental timing with a major UK holiday resulting in high flight costs). The overspend in travel was compensated by lower than budgeted personnel costs, maintaining the overall budget and levels of effort and delivery.
Appendix I

Report on the consultative workshop
A MEMORY OF ELEPHANTS

CONSULTATIVE WORKSHOP REPORT

12 June 2014

University of Oxford e-Research Centre
Consultative Workshop Report

Summary
From the ElEPHãT project’s inception we wanted its outputs to be useful to stakeholders inside and outside the project. To this end, we organized a consultative workshop with representatives of the scholarly community as one of the first activities of the project. We invited participants from a variety of academic backgrounds, including editors and book historians, and colleagues from libraries and publishers.

The day was arranged to exchange knowledge. Kevin Page, Judith Siefring, and Pip Willcox introduced participants to the HathiTrust, to Early English Books Online Text Creation Partnership (EEBO-TCP), Linked Data, the Semantic Web, and some outcomes of the Sustaining the EEBO-TCP Corpus in Transition (SECT) project. We designed breakout and roundtable discussions to gather qualitative data, and a questionnaire to gather quantitative data. These data, and continuing correspondence with participants from the day, informed the decisions we made as to what to develop in the project.

The starting point for many of the participants’ scholarly questions was metadata, particularly authors, dates, and imprint information. This guided the areas of our technical development throughout the project. The workshop led to subsequent correspondence and meetings with participants, particularly following the employment of the project’s Research Associate, Terhi Nurmikko-Fuller. The workshop was a successful illustration of the importance of scholarly community building and engagement.

Project Overview
The ElEPHãT project demonstrates the use of Linked Data for combining, through worksets, information from independent collections into a coherent view which can be studied and analyzed to facilitate and improve academic investigation of the constituents.

The project has been a joint endeavour between the University of Oxford e-Research Centre, where software development was undertaken, and the Bodleian Libraries, which brought expert understanding of the collections. The project began with a consultative workshop with scholars from the field, and progressed over 9 months with fortnightly software sprints, each planned and evaluated by a Technical and Scholarly Steering Committee comprising experienced members from the e-Research Centre and Bodleian.

The project produced software to expose the necessary metadata from the individual collections so that it can be used to parameterize aggregate worksets, in accordance with the model being developed as part of the primary WCSA award. This presented a number of challenges both in the structuring and alignment of the metadata internally and with external resources, and the development of software to investigate and capture alignment, links, and analysis, which the project also developed.

The technical challenges were motivated and scoped through focus on a particular area of academic research: the study of early English books (1473-1700). This is a field in which the project team has significant familiarity and expertise.

The prototype built integrated worksets drawing resources from the HathiTrust and from the Early English Books Online Text Creation Partnership (EEBO-TCP) collection, which is co-led by the Bodleian Libraries and the University of Michigan Library. EEBO-TCP provides an interesting and complementary corpus to HathiTrust, being focussed upon high quality images and accurate transcriptions of items which are often found in libraries’ special collections. Preliminary investigation showed a significant intersection between the collections, indicating fruitful potential avenues for investigation of alignment and linking through worksets, including editions that span the collections, and networks of common authors, publishers, and locations.
Area of Scholarly Study: Early English Print

We elected to focus our prototyping work on the study of, and in relation to, early English books, from the beginning of print in English (1473) to 1700. While there were many generic technical challenges to investigate and overcome in developing a workset demonstrator, scoping our work to early English print brought a number of specific benefits, while enabling us to illustrate the generalizable benefits of our approach, particularly the use of Linked Data.

Motivation

The technical challenges of the project are best viewed through the lens of use cases motivated by actual scholarly questions. Such questions enabled the scoping and prioritizing of technical requirements according to needs which can be trialled, evaluated, and refined in an iterative and incremental manner—as opposed to a monolithic, predetermined design. To develop these questions, we needed to consult researchers, students, and librarians who use these texts “in the wild”.

Identifying workset problems from a scholarly perspective in turn helped identify the metadata required to parameterize that workset. This helped prioritize the elements of RDF encoding and alignment between collections and external entities, which enabled effective use of these parameters.

Participants

We were delighted to benefit from the expertise of colleagues from eight institutions, and from a range of academic and professional backgrounds, including book historians, crowd-sourcing and citizen science experts, digital curators, digital humanists, early modern historians, editors, literary scholars, special collections librarians.

- Jacqueline Baker, Oxford University Press
- James Baker, British Library
- Rebecca Bullard, University of Reading
- Stephen Gregg, Bath Spa University
- Ben Higgins, University of Oxford
- Rupert Mann, Oxford University Press
- Jane Potter, Oxford Brookes University
- Kirsty Rolfe, University College London
- Matthew Symonds, University College London
- Victoria van Hyning, University of Oxford
- Sarah Wheale, Bodleian Libraries
- Elizabeth Williamson, University of Oxford

Pip Willcox (ElEPHãT Co-investigator) led the workshop. Kevin Page (ElEPHãT Principal Investigator), and David De Roure and Judith Siefring, from the project’s Technical and Scholarly Steering Committee, presented talks and, with Michael Popham, took an active part helping to facilitate the workshop.

Workshop Programme

Following refreshments, kindly provided by the Oxford e-Research Centre, and introductions, Kevin Page introduced participants to Linked Data and the idea of worksets (‘The ElEPHãT from above’), and Pip Willcox gave an overview of the HathiTrust and EEBO-TCP corpora (Feeding the ElEPHãT). Kevin presented the work of the project (The ElEPHãT enclosure), and Judith Siefring, from the project’s Technical and Scholarly Steering Committee, presented some findings of the SECT project, offering examples of the types of scholarly questions the project could address (What the ElEPHãT shouldn’t forget). The presentations that accompanied these talks appear as appendices to this report, as does the handout prepared for participants.

After lunch, where informal discussions continued, David De Roure gave a wider context to the ElEPHãT project, describing other projects across the University of Oxford, particularly at the Oxford e-Research Centre (The ElEPHãT from University of Oxford e-Research Centre Bodleian Libraries
a distance). Judith Siefring and Pip Willcox facilitated breakout discussion groups (Herding ElEPHãTs) with notes taken by Michael Popham and David De Roure respectively, before more refreshments, and a roundtable discussion (Talking ElEPHãTs). Participants were asked to complete a questionnaire (see appendix) before the day ended.

**Results**

Comments were collected from breakout groups and reported back and further discussed during the roundtable discussion. An outline of the discussions with results of the questionnaire appear below.

Discussion was both broad-ranging and detailed, as participants exchanged ideas and generated new possibilities and directions for scholarly investigation. The researchers’ varied perspectives provided insight into the potential they saw for Linked Data, allowing us to take ideas forward directly from the community, translating these into use cases.

**Recommendations**

Of paramount importance to all areas of research was the discoverability of texts, through searching accurate metadata, and through less immediately obvious links, such as networks of printers and authors. For this prototype project, then, it was clear that use cases concentrating on imprint information—authors, dates, and places of publication—was the natural starting point.

Information gathered from information in the imprint or colophon is augmented in EEBO-TCP and the HathiTrust by the research published in Pollard and Redgrave’s and Wing’s Short Title Catalogues, the Thomason Tracts, and the Early English Books Tract Supplement. Additional metadata includes genre (for example, “biography”), and the ability to search by explicit as well as inferred metadata is important.

While participants were content to be guided through example use cases, it was unanimously agreed that in order to use resources to produce worksets themselves, human-friendly interfaces were vital.

**Future Work**

Many ideas for future work came out of the workshop’s discussions, formal and informal, as well as a scholarly community willing to continue advising on any future work.

Addressing further enrichment and annotation of metadata content is beyond the scope of current work. The possibility of coordinating HathiTrust and EEBO-TCP metadata with scholarship as it develops, including through book trades indexes, will be facilitated by describing it through Linked Data. EEBO-TCP title pages are marked up separately, though no richly, in the TEI XML: this could be developed, including programmatically. The physicality of the digital could assist with this, for example using OCR to map the spaces on a page.

Networks are a rich seam of research which Linked Data can help to trace and be used to visualize. These include networks of people—authors, printers, booksellers, bookbinders—and works, as well as more abstract abstract concepts, such as the movement of ideas and language.

When Linked Data is applied to HathiTrust and EEBO-TCP full texts, the potential will be extraordinary. Useful worksets could be created around occurrence of particular terms, such as “Whig” and “Tory”. Work could be undertaken on flows of ideas, as well as piracy, plagiarism, and co-citation (actual or falsified). Linking data across languages would enrich this still further. Linking metadata and full-texts with historical databases and encyclopaedias would create a gold standard of research resource.

Pip Willcox

Bodleian Libraries, University of Oxford

Oxford, 15 June 2015
Appendix 1: Notes from the Discussions

Notes from the breakout groups were taken by David De Roure and Michael Popham. Additional notes were taken during the roundtable discussion by Pip Willcox, and have been collated and re-ordered here.

A) Stationers’ functions
Distinguishing between trade functions of members of the Stationers’ Company, for example people who print, bind, and sell books, is currently difficult in EEBO.

B) Prioritizing text or container?
EEBO-TCP silently prioritizes text over its container (typically a book). HathiTrust leans towards prioritizing container over text.

C) Releasing data without releasing content
Releasing data about texts is useful, even where content is restricted, for example, pages per book, or illustrations per page.

D) Seeing through the eyes of a machine
When a human views a book, the heft, thickness of paper, weight and size of a book are taken in. Combining this information with physical information more easily provided by a machine will be a powerful tool, potentially leading to new insights about book production.

E) Comparing texts
HathiTrust content is most useful when there is more than one version of a text, or where the output of a printer can be considered together.

F) Variant spelling
Comparing modern and original editions of a book might be used to overcome the difficulties of variant spelling.

G) Recognizing anonymous and pseudonymous publications
Anonymous and pseudonymous publication are commonplace in this period. Programatically identifying features or style, for example, meter and form, could help identify individuals or groups. Scepticism was expressed regarding how much machines can be taught to recognize, and whether such results would be accepted by the academic community (cf. authorship debates).

H) Recognizing genres
Teaching a machine to recognize genre would necessitate revisiting implicit scholarly assumptions. Could texts be marked up semantically to facilitate this?

I) False or absent metadata
Where metadata is derived from deliberately falsified title pages or colophons, can it be annotated as such, then used to cross-search other occurrences or co-occurrences of the same information? Similar use could be made of metadata derived from extra-textual sources.

J) Understanding limitations
This data will enable scholars to answer new questions, but it has limits: these must be clearly stated. Where feasible, resources could be linked to each other to provide more data.
K) Human limitations of metadata
When we search, we find what the cataloguer was interested in. Resources are one filter, but another is the cataloguer's interests, and individual library policy and practice.

L) Re-considering working practices
“What kinds of new questions can I generate?” For example, a computer can count numbers and letters but not see meaning, for example irony. Researchers need to learn new working protocols.

M) Tacit knowledge
Working with, and particularly teaching, digital methods forces tacit knowledge to be made explicit.

N) Multidisciplinarity
Economic historians bring different tacit knowledge from literary scholars, and different again from computer scientists. Working collaboratively to create, research, teach, and learn with digital resources will seed knowledge across and between disciplines.

O) Funding
Being forced to articulate methods and methodologies has an unintended outcome of improving research funding applications.

P) Using whitespace
Where OCR text is not of high quality (as is frequent with early modern text), can whitespace be used to identify printing practice or to match copies of the same work?

Q) Cost and marketing
Number of pages, illustrations, paper size, amount of whitespace could be used to infer the cost of publications. Similarly, blank pages (for example, in almanacs) could be used to find publications where the reader could interact with the text.

R) Correcting text
Could corrections be made programmatically? Remembering there are errors in human-transcribed texts too, could potential errors be highlighted for machine or human correction?

S) Contextual corrections
Could machines be taught to correct contextually, cf. predictive text.

T) Supporting scholarly editing
Accessing different states and editions of a text is of primary importance to editors. Could these be programmatically brought together, for human consideration and editing?

U) Errata and reception
Combining printed lists of errata with metadata about ownership, provenance, and annotation, could inform understanding of textual reception.

V) Date of publication
Particularly during times of war or political contention, month as well as year of publication is of interest. Could this be added to metadata where it is known or can be derived from the text?
W) Networks
Uncovering networks of printers, writers, and booksellers is useful. Could intellectual and temporal networks also be mapped, e.g. publications around events or contentious topics. These might be found by tracing phrases and other information on title pages, as well as through close textual reading.

X) Contentious words
Co-occurrence of particular words and phrases would make an interesting workset, for example “Whig” and “Tory”.

Y) International, multilingual linking
The growing interest in translation studies could be supported by linking encoded texts with collections in other languages, for example in the HathiTrust, or the Medici Archive. Tracing networks of people and ideas across languages is of increasing interest to scholars.

Z) Linking across genres of resources
Linking data between disparate genres of resources, for example HathiTrust and EEBO-TCP with the Oxford Dictionary of National Biography, would provide rich research, between authors, networks, style, language, and ideas.
Appendix 2: Questionnaire Data

We received eight completed questionnaires from the workshop, whose results appear below.

<table>
<thead>
<tr>
<th>Question</th>
<th>Very often</th>
<th>Often</th>
<th>Occasionally</th>
<th>Never</th>
<th>No ans</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>How often do you use EEBO-TCP material in your own work?</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>How often do you use HathiTrust materials?</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>How often do you search digital collections?</td>
<td>7</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>How useful and reliable do you find digital collections?</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>Would you find it useful to be able to search across the EEBO-TCP and HathiTrust materials together?</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>Based on what you have learned today, would you be interested in the research outputs of the Eiephit project?</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8</td>
</tr>
</tbody>
</table>